HETEROGENOUS GRAPH NEURAL NETWORKS FOR SPECIES DISTRIBUTION MODELING

Lauren Harrell[a][1]**, Christine Kaeser-Chen**[b]**,**
Burcu Karagol Ayan[b]**, Keith Anderson**[b]**, Michelangelo Conserva**[a]**,**
Elise Kleeman[a]**, Maxim Neumann**[b]**, Matt Overlan**[b]**, Melissa Chapman**[a]**,**
Drew Purves[b]

[a]Google Research [b]Google DeepMind
[1]Corresponding author: laurenharrell@google.com

ABSTRACT

Species distribution models (SDMs) are necessary for measuring and predicting occurrences and habitat suitability of species and their relationship with environmental factors. We introduce a novel presence-only SDM with graph neural networks (GNN). In our model, species and locations are treated as two distinct node sets, and the learning task is predicting detection records as the edges that connect locations to species. Using GNN for SDM allows us to model fine-grained interactions between species and the environment. We evaluate the potential of this methodology on the six-region dataset compiled by National Center for Ecological Analysis and Synthesis (NCEAS) for benchmarking SDMs. For each of the regions, the heterogeneous GNN model is comparable to or outperforms previously-benchmarked single-species SDMs as well as a feed-forward neural network baseline model.

1 INTRODUCTION

Understanding where species are distributed across the globe is essential not only for advancing our scientific understanding of ecological processes (Guisan & Rahbek, 2011), but also to enable the strategic implementation of conservation measures and policies (Guisan et al., 2013). Species Distribution Models (SDMs) provide a means for predicting species occurrence across landscape scales, generally leveraging species occurrence data and environmental covariates. While a multitude of methods have been developed and explored to predict species distributions (Elith & Leathwick, 2009; Johnston et al., 2020; Anderson, 2012), many challenges persist. First, data on species is often presence-only (only provides information about where a species is found, not where it is absent), making it difficult to identify the environmental conditions that define the boundaries of a species range. Second, underlying biodiversity data used to train SDMs is often biased (Chapman et al., 2024; Hughes et al., 2021), risking that models reflect these data disparities or observational processes, rather than underlying ecological processes intended (A. Lee-Yaw et al., 2022). Third, SDMs are difficult to evaluate at scale, due to the lack of "ground-truth" data (Beery et al., 2021).

A number of recent works have focused on addressing the challenge of utilizing presence-only records for SDMs (Cole et al., 2023; Valavi et al., 2022). In this paper, we focus on learning the interaction between the environment and species, the first step towards understanding ecological processes. We present a novel approach to SDMs by using heterogeneous graph neural networks (GNN). We model locations and species as graph nodes, and their interactions as explicit graph edges in our GNN. This enables us to learn the representations of location, species, and their interactions, supervised by the species occurrences records. The GNN model is extensible to different observation and species data modalities and can model various ecological interactions as distinct edge types. Our proof-of-concept results on a benchmark dataset are comparable with common SDM approaches as well as an alternative deep-learning approach.

2 THE CASE FOR GRAPH NEURAL NETWORKS FOR SPECIES DISTRIBUTIONS

Graph Neural Networks (GNNs) are a class of deep learning models that capture relationships between entities, by processing data structured as graphs, where nodes represent entities, and edges represent their connections (Scarselli et al., 2009). Broadly speaking, a graph $\mathcal{G}$ is defined as a set of nodes and edges, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the set of nodes and $\mathcal{E}$ the set of edges. Edges are directed when there is an ordering or dependency that is not symmetric between two nodes and undirected when the relationship between two nodes is the same in both directions. A key feature of GNNs is the ability to pass messages between nodes through the edge sets, resulting in node representations that aggregate information from their neighboring nodes.

Heterogeneous graphs can contain multiple types of nodes and edges, as opposed to a single node and edge type in homogeneous graphs. Representing data through heterogeneous graphs unlocks flexibility in data structures and ability to learn through multiple types of relationships. Recent development of deep-learning approaches to SDM problems treat the species detection/non-detection or presence/absence as a label set (Cole et al., 2023; Estopinan et al., 2022; Gillespie et al., 2024), which limits the information that may be relevant from species-specific features such as taxonomic hierarchy, size, mass, or trophic relationships. Heterogeneous graphs, on the other hand, can represent a variety of data without requiring all observations being represented as feature vectors with the same length, allowing the flexible incorporation of information into the model such as species traits across different taxonomic groups and complex remote-sensing feature data for locations. We can also represent and learn from multiple types of relationships, such as taxonomic similarity and location proximity, to better represent the ecological systems.

We represent the observed species detection records as a heterogeneous graph $\mathcal{G}(\mathcal{V}^S, \mathcal{V}^L, \mathcal{E}^{L2S}_{\text{Detection}}, \mathcal{E}^{S2L}_{\text{Detection}})$ with distinct node sets for species $\mathcal{V}^S$, and locations $\mathcal{V}^L$. We represent detections as edges ($\mathcal{E}^{L2S}_{\text{Detection}}, \mathcal{E}^{S2L}_{\text{Detection}}$) that connect species to location nodes in both directions. For location $v_i^L \in \mathcal{V}^L$ and species $v_j^S \in \mathcal{V}^S$, we include edges $e_{i,j}^{L2S}$ and $e_{j,i}^{S2L}$ in the edge set if species j is detected at location i as shown in Figure 1. Our learning task is a link-prediction task (Zhang & Chen, 2018) to infer whether an unseen "detection" edge $e_{i',j'}^{L2S}$ should exist between location node $v_{i'}^L$ and species node $v_{j'}^S$. One way to compute the probability of $e_{i',j'}^{L2S}$ is to decode from learned representations of $v_{i'}^L$ and $v_{j'}^S$, where the species representation is computed by aggregating information from the location nodes with known detections through $\mathcal{E}^{L2S}_{\text{Detection}}$. More details can be found in Appendix A.2.1. The training graph may also include a separate non-detection edge set to enable message passing between species and locations without observed detections, such that separate model weights can be learned for aggregating information along these edges.

3 APPLICATION OF GNNs TO THE NCEAS BENCHMARKS

We apply the GNN approach to the benchmarks from Elith et al. (2020) (which we will refer to as "NCEAS benchmarks"). The data is comprised of surveys from six distinct regions[1], where presence-only (PO) records are available for training, and presence-absence (PA) records are used for testing. We select this dataset because it has been used to benchmark existing presence-only SDMs (Valavi et al., 2022) and is publicly available.

For each of the six regions, the dataset provides between 11 and 13 pre-extracted environmental features, in addition to the spatial coordinate information for each data point. The goal is to predict the presence-absence status of between 20 and 54 species, depending on the region, given the presence-only detections in the training data. The characteristics of the training and test data for each region can be found in the Appendix in Table 2. The dataset also supplies 10,000 sampled background locations per region as part of the training data.

---

[1]The regions are: (1) Ontario, Canada (CAN), (2) South American countries (SA), (3) Switzerland (SWI), (4) New South Wales (NSW), (5) Australian Wet Tropics (AWT), and (6) New Zealand (NZ).
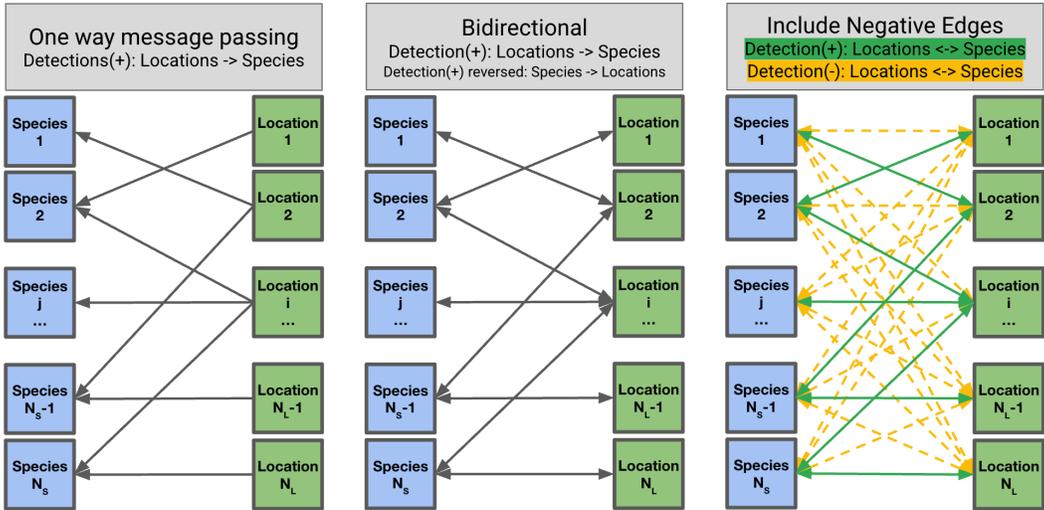
Figure 1: Example of bipartite heterogeneous graph structure where species nodes are connected to location nodes through detection edges. The first graph on the left shows one way message passing from locations $v_i^L \in \mathcal{V}^L$ to species $v_j^S \in \mathcal{V}^S$ through detections $\mathcal{E}_{\text{Detection}}^{L2S}$. The middle bipartite graph includes the reversed edge set $\mathcal{E}_{\text{Detection}}^{S2L}$ that sends information from species nodes back to location nodes. The graph on the right includes message passing through (pseudo)-negative edges as a distinct edge set.

3.1 GRAPH DATA REPRESENTATION

We build a heterogeneous graph with two different node types to represent the data. One node set represents the locations, with provided environmental attributes as features. The other node set represents the species. Features for each species node include the species ID encoded as a one-hot vector, and optionally the group information (e.g., plant or bird), for two of the NCEAS regions (NSW and AWT) where species group is provided (Figure 3 in Appendix). We create an edge set for detections between each location-species pair where a record exists in the PO data. For locations with multiple detections and potential variation in the environmental features, we aggregate the location features using the mean. Species distribution is then learned through an link-prediction task on the graph for whether a detection edge should exist between a location node and a species node.

3.2 MODEL TRAINING

We train separate GNN models for each region. During training, we train on the graph built from the training PO data with the background points appended as additional location nodes. Message passing on the known detection edges can be modeled as one-way from location to species, or bi-directional where there is also message passing from species nodes back to location nodes. We report results on both model types below.

Species detection at locations is formulated as a link prediction task on this graph. For each target $< v_{i'}^L, v_{j'}^S >$ node pair, we compute a score based on the dot product of the output latent features for the species and location nodes after message passing steps, and apply a sigmoid function on the score to compute the probability of the target link.

At each epoch during training, we sample training examples of target links with both positive (i.e., known detections) and negative labels. Like most presence-only SDMs (Cole et al., 2023; Zbinden et al., 2024; Barbet-Massin et al., 2012), we construct pseudo-negative data by assuming unobserved species are not present in PO locations and all species are not present in background locations. In our experiment, we can utilize the full set of positive examples from the PO dataset in each epoch and also include sampled pseudo-negative examples. The ratio of pseudo-negative examples to positive examples is controllable by a hyperparameter. Training loss is computed as the cross-entropy between link probability and the binary target link labels.

3

Our GNN model is based on the Interaction Network (IN) architecture (Battaglia et al., 2016). In addition to the IN message passing networks ϕ_O and ϕ_R, we also add separate node set and edge set encoders and decoders before and after the message passing steps to facilitating feature learning. All neural network components are modeled with multi-layer perceptrons (MLPs). We ablate the hidden layer size, the number of hidden layers for these MLPs, and also the number of message passing steps in our experiments; details can be found in Appendix A.3.

3.3 EVALUATION

We adopt the area under the receiver operating characteristic curve (AUC_{ROC}) metric for evaluation, similar to Valavi et al. (2022). For each region, we report the average AUC_{ROC} across all species. When evaluating the GNN, we add the test location nodes to the training graph, but do not add any edges connecting them to species, as the detection status is unknown. We run the trained GNN with message passing on this test graph. Without any edges, the test location nodes' latent features are essentially computed by the learned location encoder, while the species nodes still get message passing from the training graph. We then compute the dot product on features between test location nodes and each of the species nodes, and apply the sigmoid function to compute the link probabilities.

3.4 BASELINE METHOD

To study the effects of applying the graph structure to SDM, we compare with a feed-forward MLP model as the baseline. The MLP model uses the environmental features from the NCEAS dataset as input, and has a multi-class classification head as the output. To train the model, we apply the same pseudo-negative generation strategy as above: all observed species classes at a given location are assigned positive labels, while all other species are assigned negative labels. We also assign all species negative labels for background locations. The training batches are composed by the presence-only locations and background locations, where the mixing ratio is controlled by a hyperparameter. In our implementation, we normalize all input features to range [-1, 1] and apply a small random noise to each input feature as perturbation during training. The model is trained with the sigmoid cross entropy loss on the output logits.

We ablate the hidden layer size, the number of hidden layers, and the ratio of sampled negative data per training batch. In the following, we report results on models with hidden layer size of 32 and 4 layers, except for the NSW region in the NCEAS benchmark where a 6-layer MLP performs better. We observe that, contrary to previous findings, mixing pseudo-negative examples from *background* locations (where we assume all species are absent) in training has no benefit to model performance, as long as the *PO* locations have pseudo-negative labels assigned for unobserved species. We plan to study the effect of background sampling more closely in a future work.

3.5 RESULTS

A summary of the top AUC_{ROC} models by model type, including GNNs, the baseline MLP, and prior results from Valavi et al. (2022) are displayed in Figure 2. Some key observations are:

1. Compared to prior results, GNN has better performance on 3 regions (SWI, CAN, AWT) and comparable performance on 2 regions (NZ, NSW), with a relative increase of 23.5% over prior methods in CAN. Only on the SA region does the GNN have lower performance. We suspect this is due to the SA dataset being particularly sensitive to over-fitting, which is also observed when training the baseline MLP on this region, in addition to the relatively low number of unique locations in the test data (152 total).

2. Compared to baseline MLP models, GNN is significantly better on AWT, comparable on 3 regions (NZ, NSW, CAN), and slightly worse on 2 regions (SWI, SA). The performance gain on AWT is particularly promising, as this region does have two taxonomic groups and thus species features beyond the species ID; however the additional species groups in NSW did not produce substantial performance differences.

3. Among the GNN models, we observe variation across regions in the performance trade-off between one-way GNN and bidirectional GNN message passing. Better understanding on the effects of edge directions and graph complexity will be explored in future work.
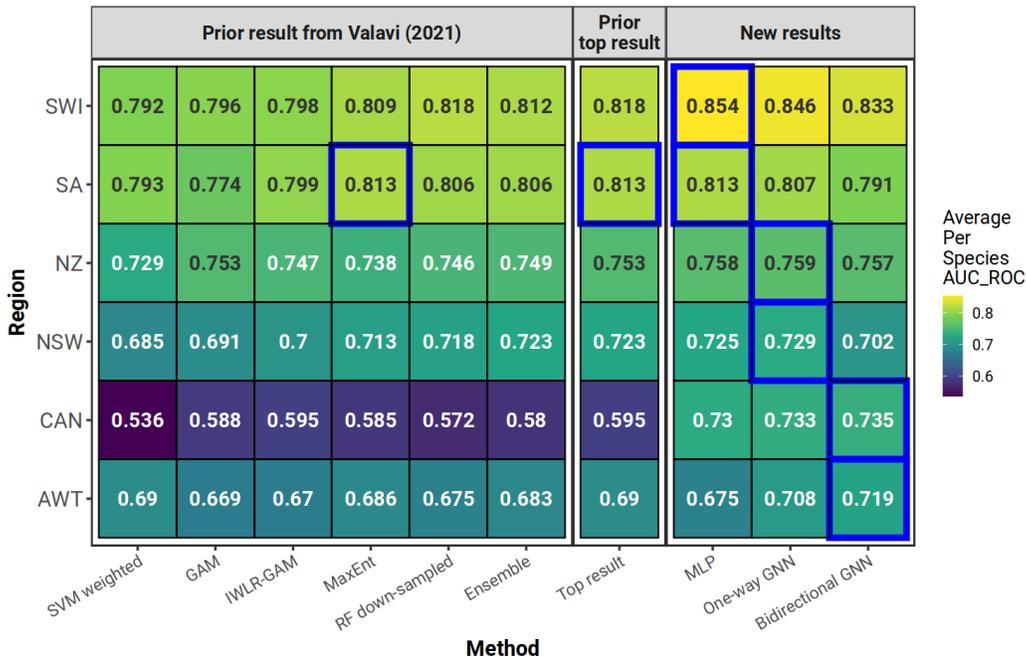
Figure 2: AUC$_{ROC}$ averaged across species per site by region and model methodology. The values of prior results were taken from the top scoring models in Valavi et al. (2022). Top result per region highlighted in blue.

4. As a caveat, prior results in Valavi et al. (2022) are all trained as single-species models, and thus do not leverage the potential information shared between species. Both the GNN methods and our baseline MLP method are trained as multi-species models, which likely contributes partly to the observed performance improvements.

4 DISCUSSION AND FUTURE WORK

In this work, we provide a basic example of using heterogeneous GNNs for species distribution modeling through a bipartite graph as a proof-of-concept, and show that performance is on par with or exceeds other SDM methods for presence-only data. We are continuing to develop the approach further for more complex, feature-rich data that includes blending presence-only with presence-absence data for model training and graph structures, more informative species traits, and remote-sensing-derived environmental features. We are also developming improved graph-based learning techniques such as including message-passing through additional edges from species-to-species and location-to-location and weighted loss functions for accounting for different pseudo-negative strategies.

5 ACKNOWLEDGEMENTS

REFERENCES

Julie A. Lee-Yaw, Jenny L. McCune, Samuel Pironon, and Seema N. Sheth. Species distribution models rarely predict the biology of real populations. *Ecography*, 2022(6):e05877, 2022.

Robert P Anderson. Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. *Annals of the New York Academy of Sciences*, 1260(1):66–80, 2012.

Morgane Barbet-Massin, Frédéric Jiguet, Cécile Hélène Albert, and Wilfried Thuiller. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods Ecol. Evol.*, 3(2):327–338, April 2012. doi: 10.1111/j.2041-210x.2011.00172.x.

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv*, June 2018.

Sara Beery, Elijah Cole, Joseph Parker, Pietro Perona, and Kevin Winner. Species distribution modeling for machine learning practitioners: A review. *arXiv [cs.LG]*, 3 July 2021.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/jax-ml/jax`.

Melissa Chapman, Benjamin R Goldstein, Christopher J Schell, Justin S Brashares, Neil H Carter, Diego Ellis-Soto, Hilary Oliva Faxon, Jenny E Goldstein, Benjamin S Halpern, Joycelyn Longdon, et al. Biodiversity monitoring for a just planetary future. *Science*, 383(6678):34–36, 2024.

Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, and Oisin Mac Aodha. Spatial implicit neural representations for global-scale species mapping. *arXiv [cs.LG]*, 5 June 2023.

Jane Elith and John R Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 40(1):677–697, 2009.

Jane Elith, Steven J Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J Yates. A statistical explanation of maxent for ecologists. *Diversity and distributions*, 17(1):43–57, 2011.

Jane Elith, Catherine Graham, Roozbeh Valavi, Meinrad Abegg, Caroline Bruce, Andrew Ford, Antoine Guisan, Robert J Hijmans, Falk Huettmann, Lucia Lohmann, Bette Loiselle, Craig Moritz, Jake Overton, A Townsend Peterson, Steven Phillips, Karen Richardson, Stephen Williams, Susan K Wiser, Thomas Wohlgemuth, and Niklaus E Zimmermann. Presence-only and presence-absence data for comparing species distribution modeling methods. *Biodivers. Inf.*, 15(2):69–80, July 2020. doi: 10.17161/bi.v15i2.13384.

Joaquim Estopinan, Maximilien Servajean, Pierre Bonnet, François Munoz, and Alexis Joly. Deep species distribution modeling from sentinel-2 image time-series: A global scale analysis on the orchid family. *Front. Plant Sci.*, 13:839327, 22 April 2022. doi: 10.3389/fpls.2022.839327.

GBIF: The Global Biodiversity Information Facility. What is gbif?, 2024. URL `Availablefromhttps://www.gbif.org/what-is-gbif`.

Lauren E Gillespie, Megan Ruffley, and Moises Exposito-Alonso. Deep learning models map rapid plant species changes from citizen science and remote sensing data. *Proceedings of the National Academy of Sciences*, 121(37):e2318296121, 2024.

Jonathan Godwin*, Thomas Keck*, Peter Battaglia, Victor Bapst, Thomas Kipf, Yujia Li, Kimberly Stachenfeld, Petar Veličković, and Alvaro Sanchez-Gonzalez. Jraph: A library for graph neural networks in JAX., 2020. URL `http://github.com/deepmind/jraph`.

Antoine Guisan and Carsten Rahbek. Sesam–a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, 38(8):1433–1444, 2011.

Antoine Guisan, Reid Tingley, John B Baumgartner, Ilona Naujokaitis-Lewis, Patricia R Sutcliffe, Ayesha IT Tulloch, Tracey J Regan, Lluis Brotons, Eve McDonald-Madden, Chrystal Mantyka-Pringle, et al. Predicting species distributions for conservation decisions. *Ecology letters*, 16(12): 1424–1435, 2013.

Alice C Hughes, Michael C Orr, Keping Ma, Mark J Costello, John Waller, Pieter Provoost, Qinmin Yang, Chaodong Zhu, and Huijie Qiao. Sampling biases shape our view of the natural world. *Ecography*, 44(9):1259–1269, September 2021. doi: 10.1111/ecog.05926.

Alison Johnston, Nick Moran, Andy Musgrove, Daniel Fink, and Stephen R Baillie. Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, 422: 108927, 2020.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Alexander Pritzel, Suman Ravuri, Timo Ewalds, Ferran Alet, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Jacklynn Stott, Oriol Vinyals, Shakir Mohamed, and Peter Battaglia. Graphcast: Learning skillful medium-range global weather forecasting. *arXiv*, 2022.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Steven J Phillips, Robert P Anderson, and Robert E Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259, 2006.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605.

Roozbeh Valavi, Gurutzeta Guillera-Arroita, José J Lahoz-Monfort, and Jane Elith. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecol. Monogr.*, 92(1), February 2022. doi: 10.1002/ecm.1486.

Robin Zbinden, Nina van Tiel, Benjamin Kellenberger, Lloyd Hughes, and Devis Tuia. On the selection and effectiveness of pseudo-absences for species distribution modeling with deep learning. *Ecological Informatics*, 81:102623, July 2024. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2024. 102623. URL `http://dx.doi.org/10.1016/j.ecoinf.2024.102623`.

Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. *Advances in neural information processing systems*, 31, 2018.

A   APPENDIX

A.1   SOFTWARE

Our models were trained using a modified version of the jraph (Godwin* et al., 2020) software package that was designed for training homogeneous graph neural networks with JAX (Bradbury et al., 2018), an open-source Python library. An extension of jraph called TypedGraphs was developed for heterogeneous graphs, and deep learning on these graphs was implemented and open-sourced via GraphCast, which developed a GNN for global weather forecasting (Lam et al., 2023; 2022).

A.2 MODEL AND TRAINING METHODOLOGY DETAILS

Most SDM methodologies that leverage presence-only data are ultimately learning about the location feature distribution of the locations where the species has been observed. For example, the output of the popular SDM method MaxEnt (Phillips et al., 2006) is effectively a ratio of the joint distribution of the environmental features from presence-only locations to the joint distribution of the background features (Elith et al., 2011). Message passing in the GNN is a different way to learn representations of both the species and the locations where, depending on the direction of message passing, species node sets are embedded with aggregating information from the locations and vice versa. For detection edges that go from from locations to species ($\mathcal{E}_{\text{Detection}}^{L2S}$), the species embeddings are updated by aggregating information from locations in which there are detections of that species. If message passing occurs in the reverse direction in $\mathcal{E}_{\text{Detection}}^{S2L}$, the location embeddings are then updated with information aggregated from the species set.

A.2.1 GNN MODEL STEPS

Our GNN approach follows the *encode, process, decode* framework established by Battaglia et al. (2018). Underneath the hood, the encoder models we use in this paper are a multi-layer perceptrons that embed the species node set, location node set, and detection edge set in the same dimensional latent space Ω^d, which is a vector of length d for each node or edge. After encoding, we run a fixed number of message passing steps so that feature information is aggregated and shared through the edges from sender to receiver nodes, following the Interaction Network architecture described in Battaglia et al. (2016) and Lam et al. (2022).

Our model operates on a heterogeneous graph $\mathcal{G}(\mathcal{V}^S, \mathcal{V}^L, \mathcal{E}_{\text{Detection}}^{L2S})$, which can be expanded to include additional edge sets for message passing to include $\mathcal{E}_{\text{Detection}}^{S2L}$ for bidirectional message passing through detection edges and/or $\mathcal{E}_{\text{Non-Detection}}^{L2S}$, $\mathcal{E}_{\text{Non-Detection}}^{S2L}$ for message passing for species/location pairs without detections (i.e., "negative" edges). $\mathcal{V}^S$ represents the node sets indexing the set of species nodes v_j^S and the associated feature set for each species j. $\mathcal{V}^L$ is the node set that indexes locations, with a node v_i^L for each location i containing the set of environmental features. $\mathcal{E}_{\text{Detection}}^{L2S}$ is the edge set that connects location nodes $\mathcal{V}^L$ to species nodes $\mathcal{V}^S$ where there exists detections, and $\mathcal{E}_{\text{Detection}}^{S2L}$ is the edge set that reverses the direction of the detection edges to share information from species $\mathcal{V}^S$ to locations $\mathcal{V}^L$. If included in the model specification, we may also include "negative" detection edges as separate edge sets to aggregate information from location/species pairs without detections $\mathcal{E}_{\text{Non-Detection}}^{L2S}$ and the corresponding reverse direction for bidirectional message passing $\mathcal{E}_{\text{Non-Detection}}^{S2L}$.

Step 1: Encode features. The first step in the GNN embeds the features of all nodes and edges into the same-dimensional latent space.

$$
\mathbf{v}_j^S = \text{MLP}_{\mathcal{V}^S}^{\text{embedder}}(\mathbf{v}_j^{S,\text{features}})
$$

$$
\mathbf{v}_i^L = \text{MLP}_{\mathcal{V}^L}^{\text{embedder}}(\mathbf{v}_i^{L,\text{features}})
$$

$$
\mathbf{e}_{\mathcal{V}^L \to \mathcal{V}^S}^{L2S,\text{Detection}} = \text{MLP}_{\mathcal{E}_{\text{Detection}}^{L2S}}^{\text{embedder}}(\mathbf{e}_{\mathcal{V}^L \to \mathcal{V}^S}^{L2S,\text{Detection,features}})
$$

$$
\mathbf{e}_{\mathcal{V}^S \to \mathcal{V}^L}^{S2L,\text{Detection}} = \text{MLP}_{\mathcal{E}_{\text{Detection}}^{S2L}}^{\text{embedder}}(\mathbf{e}_{\mathcal{V}^S \to \mathcal{V}^L}^{S2L,\text{Detection,features}}) \quad (optional)
$$

$$
\mathbf{e}_{\mathcal{V}^L \to \mathcal{V}^S}^{L2S,\text{Non-Detection}} = \text{MLP}_{\mathcal{E}_{\text{Non-Detection}}^{L2S}}^{\text{embedder}}(\mathbf{e}_{\mathcal{V}^L \to \mathcal{V}^S}^{L2S,\text{Non-Detection,features}}) \quad (optional)
$$

$$
\mathbf{e}_{\mathcal{V}^S \to \mathcal{V}^L}^{S2L,\text{Non-Detection}} = \text{MLP}_{\mathcal{E}_{\text{Non-Detection}}^{S2L}}^{\text{embedder}}(\mathbf{e}_{\mathcal{V}^S \to \mathcal{V}^L}^{S2L,\text{Non-Detection,features}}) \quad (optional)
$$

Step 2: Processing/message passing. In the next step, our model conducts message passing and node/edge embedding updates using the aggregated features. For each step in the number of message passing steps, the node and edge latent feature sets are updated using the interaction network, starting with updating the edge sets based on the information of the sender/receiver node pairs. For example, for the detection edge set from locations to species ($\mathcal{E}_{\text{Detection}}^{L2S}$) included in all model versions, the update is computed as follows:

$$\mathbf{e}^{L2S,\text{Detection}'}_{\mathcal{V}^L_i \to \mathcal{V}^S_j} = \text{MLP}^{\text{processor}}_{\mathcal{E}^{L2S}_{\text{Detection}}}([\mathbf{e}^{L2S,\text{Detection}}_{v^L_i \to v^S_j}, \mathbf{v}^L_i, \mathbf{v}^S_j])$$

Note that negative edges would have similarly structured but separately learned updates, which allows for different model weights to apply to the aggregated information from the negative edges versus the positive edges. After the edge sets updates are computed, the receiver node sets (in this case the species node set) is updated by aggregating the information across the edges arriving at that node. Note that for aggregation of the edge information, if $w_k = 1$, this corresponds to the segment sum, and if $w_k = \dfrac{1}{n_e}$ we are taking the segment mean.

$$\mathbf{v}^{S'}_j = \text{MLP}^{\text{processor}}_{\mathcal{V}^S}(\mathbf{v}^S_j, \sum_{e^{L2S,\text{Detection}}_{v^L_i \to v^{S'}_j}} w_k \mathbf{e}^{L2S,\text{Detection}}_{v^L_i \to v^{S'}_j})$$

If message passing is bidirectional, then the location nodes are updated in a similar fashion as the species nodes. However, in the one-way graph, the location features do not receive messages from the species nodes and are updated independently:

$$\mathbf{v}^{L'}_i = \text{MLP}^{\text{processor}}_{\mathcal{V}^L}(\mathbf{v}^L_i)$$

The new feature values are updated with a residual connection.

$$\mathbf{e}^{L2S,\text{Detection}}_{\mathcal{V}^L_i \to \mathcal{V}^S_j} \leftarrow \mathbf{e}^{L2S,\text{Detection}}_{\mathcal{V}^L_i \to \mathcal{V}^S_j} + \mathbf{e}^{L2S,\text{Detection}}_{\mathcal{V}^L_i \to \mathcal{V}^{S'}_j}$$

$$\mathbf{v}^S_j \leftarrow \mathbf{v}^S_j + \mathbf{v}^{S'}_j$$

$$\mathbf{v}^L_i \leftarrow \mathbf{v}^L_i + \mathbf{v}^{L'}_i$$

Step 3: Decoding and output. The resulting output embeddings for the species and location nodes are encoded into a fixed size of the same dimension (defined in Table 1). In order to learn a binary classifier for whether an edge between location i and species j should exist, we need to define a function to decode to a scalar for binary classification $f(\mathbf{v}^L_i, \mathbf{v}^S_j) \to z_{L_i, S_j}$. For the purposes of this work, we use a simple dot product decoder as the similarity between the latent vectors between the species embeddings and location embeddings to decode create a single scalar to represent the score for the potential edge where $z_{L_i, S_j} = \mathbf{v}^L_i \cdot \mathbf{v}^S_j$. In the future, we could use the MLP embeddings of the edges more directly or learn alternative scoring functions that reflect the divergence of the positive versus negative edge embeddings.

A.2.2 FUTURE ARCHITECTURES

Presented in this paper is a simple proof-of-concept of leveraging heterogeneous graph neural networks where message passing occurs only on the bipartite graph between species and location node sets. In future work, we are expanding on this methodology for heterogeneous graph learning where we can separate out the encoding and processing steps of the GNN more explicitly where initial encoding occurs along the bipartite species ⇔ locations subgraph, but additional processing message passing models can be applied through location ⇔ location and/or species ⇔ species edge sets that can enable smoothing and improved representations to address sampling bias for locations or species without detections in the data.

A.2.3 LOSS FUNCTION: BINARY CROSS ENTROPY LOSS WITH LOGITS

We start with the sigmoid binary cross entropy loss with logits, which is a common loss function for link prediction tasks. For the kth edge representing species j and location i, let $score_k$ be the model output and y_k represent the true label of 1 for a detection present in the data and 0 otherwise, then for edge k, the loss is computed as:

$$prob_k = \text{sigmoid}(score_k)$$
$$loss_k = y_k * \log(prob_k) + (1 - y_k) * \log(1 - prob_k)$$

This loss function increases when scores are diverging from their expected label (i.e., high scores for "negatives" and low scores for positives). We can aggregate this loss function over all the edges using either the mean or the sum; in our examples we use the mean.

Currently our loss function does not differentiate between different pseudo-negative types such as the loss functions proposed in similar deep learning (Cole et al., 2023; Zbinden et al., 2024). It is also computed by taking the overall mean rather than averaging across loss at a species level, which may be impacting the performance on per-species AUC_{ROC} versus an overall AUC_{ROC} aggregated across all edges irrespective of species. To optimize for improved performance at the species level and to minimize effects of imbalances in the number of occurrence records per species, future work should experiment with alternative loss functions and consider optimizing per-species rather than across all edges weighted equally.

A.3 HYPERPARAMETER ABLATION EXPERIMENT SETTINGS

We experimented with a number of potential options for ablation studies to train the GNN models.

Configuration or parameter	Options
Include negative edges for message passing	[True, False]
Message passing direction	[One-way, Bidirectional]
Latent sizes	[16, 20, 32, 64]
Number of hidden layers	[1, 2, 3]
Number of message passing steps	[1, 2, 3]
Negative sampling strategy	[Uniform, Stratified ($K_{Locations}$ per species)]
Number of negative edges sampled per training step	[# of positive edges, 1000]
Proportion of negatives from PO locations (versus background)	[Random, 0.75, 1.0]
Aggregation of edges for node updates	[Segment sum, Segment mean]
Learning rates	[0.001, 0.0001]
Activation functions	[relu, leakyrelu, softplus, silu, hardsilu, sparseplus]

Table 1: Hyperparameter configuration options for GNN models

A.4 ADDITIONAL NOTES ON BENCHMARKING ON NCEAS DATA

The NCEAS data cover six distinct regions of the world where there exists disparate sets of presence-only records along with high-quality presence-absence data in the following regions:

- Ontario, Canada (CAN),
- South American countries (SA),
- Switzerland (SWI),
- New South Wales (NSW),
- Australian Wet Tropics (AWT), and
- New Zealand (NZ).

The number of species, number of location variables, unique locations with detections in the training data (excluding the additional 10,000 background locations), and unique locations in the presence-absence test data are shown in Table 2.

Region Code	Region	Species set	Number of location variables	Training Locations (with detections)	Unique Test Locations
AWT	Australian Wet Tropics, Queensland, Australia	20 birds; 20 vascular plants	13	3806	442
CAN	Ontario, Canada	30 birds	11	5063	14571
NSW	North-east New South Wales, Australia	7 bats; 8 diurnal birds; 2 nocturnal birds; 8 open-forest trees; 8 open-forest understorey vascular plants; 7 rainforest trees; 6 rainforest understory vascular plants; 8 small reptiles	13	3323	8746
NZ	New Zealand	52 vascular plants	13	3088	19120
SA	Continental Brazil, Ecuador, Colombia, Bolivia, and Peru, South America	30 vascular plants	11	1178	152
SWI	Switzerland	30 trees	13	11429	10013

Table 2: Characteristics of the NCEAS datasets for each region

A.4.1 LIMITATIONS OF NCEAS DATA

The NCEAS data provide a common set of species detection records for which different statistical and machine learning algorithms can be compared directly. Given that the majority of species distribution models are designed for use with presence-only data available for training, it is valuable to have this data that includes examples of presence-only training data alongside a held-out set of presence-absence data that can be used to directly examine how the model would perform against explicit non-detection records. With that said, it has a few important limitations:

1. Species detections were made over potentially long time frames, with no recent records after 2006. Some detections in the data occurred over 100 years ago.

2. Environmental covariates for the locations are fixed (not time-dynamic) and may not reflect the conditions at the time of detection. Alternatively, the measurements of a given detection record may not be comparable to measurements taken at other locations as counter-factuals if measured at different times.

3. Models trained/evaluated on this data assume the relationship between the species and location (habitat suitability process) remains constant over time (i.e., that species don't change their habitat niche needs and their suitable habitat would remain the same).

4. The environmental variables measured and collected may not be sufficient for predicting species presence. In practice, there will be an upper bound limit to model predictive performance if key variables are missing from the data.

5. No species-specific information - any identifying information about the species is obscured, and thus little or no species traits can be included as features.

6. Species set contained are from common taxonomic groups that are more likely to be sampled in citizen science (e.g., birds and trees) and from regions that are more likely to be represented more generally in large-scale databases like the Global Biodiversity Information Facility (GBIF: The Global Biodiversity Information Facility). Models that perform well for these species and regions may not be the best performing models for rare species or under-characterized regions.

A.5 LIMITATIONS OF BENCHMARKING COMPARISONS

The benchmarking study on the NCEAS data was limited to only methods designed for single species models (Valavi et al., 2022). Direct comparison of results of our GNN, which is trained
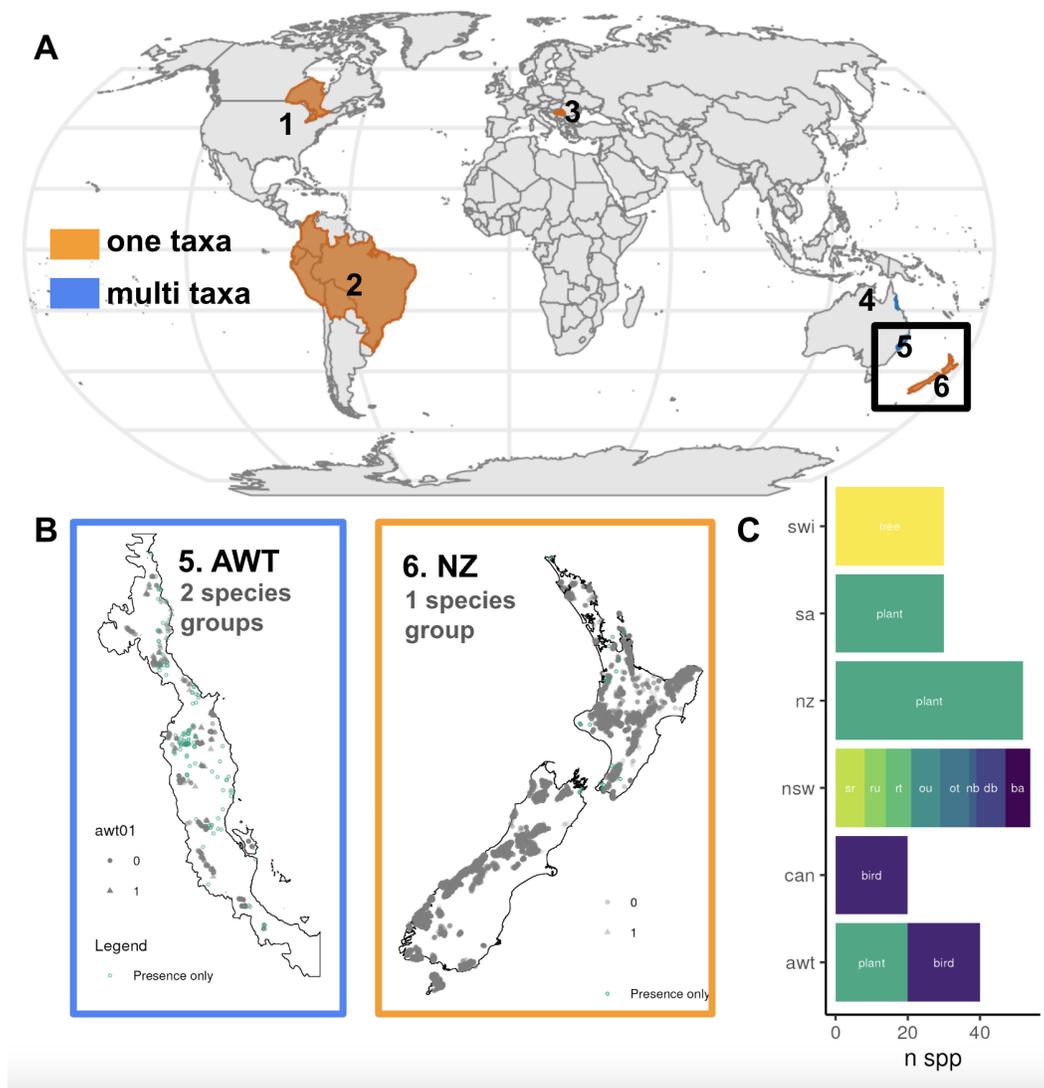
Figure 3: NCEAS Data. (A) The locations of each of the six regions around the globe shaded by number of taxonomic groups. (B) Example presence and presence/absence observations for one species in each of the SWT and NZ regions. (C) Counts of species per taxonomic group in each region.

on all available species, to models trained one species at a time have slightly different objectives and are leveraging information in different ways, and thus performance differences may not be isolated to the methodology alone but how the features and training data are processed. We are also not comparing with many other deep-learning methodologies that simultaneously estimate species distributions across multiple species, and thus we cannot show definitively that the GNN approach outperforms an alternative neural network approach. With that said, by showing our results in relation to existing benchmarking results from the single species models along with a baseline MLP model, we demonstrate the promise of this model for future applications.

Additionally, evaluating on presence-absence data is not the same as true species occupancy/non-occupancy of a given location. Absence data doesn't mean a particular species doesn't occupy a given location, just that it wasn't observed there during the survey or set of surveys included. Presence-absence data with repeated visits can be used to construct alternative types of models such as Bayesian occupancy models that aim to separate the detection process from the habitat

niche process. Using presence-absence data as a benchmarking or outcome target may still create distortions or biases in the resulting inferences, particularly for species that are much more difficult to detect even when they do occupy a given location.